Applied Soft Computing 52 (2017) 109–119

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Effects of dataset characteristics on the performance of feature selection techniques

Dijana Oreski [a,*], Stjepan Oreski [b], Bozidar Klicek [a]

[a] University of Zagreb, Faculty of Organization and Informatics, Pavlinska 2, 42000 Varazdin, Croatia
[b] Bank of Karlovac, I.G. Kovacica 1, 47000 Karlovac, Croatia

ARTICLE INFO

Article history:
Received 10 February 2016
Received in revised form 6 November 2016
Accepted 6 December 2016
Available online 19 December 2016

Keywords:
Dataset characteristics
Feature selection
Comparative analysis
Data sparsity
Feature noise

ABSTRACT

While extensive research in data mining has been devoted to developing better feature selection techniques, none of this research has examined the intrinsic relationship between dataset characteristics and a feature selection technique's performance. Thus, our research examines experimentally how dataset characteristics affect both the accuracy and the time complexity of feature selection. To evaluate the performance of various feature selection techniques on datasets of different characteristics, extensive experiments with five feature selection techniques, three types of classification algorithms, seven types of dataset characterization methods and all possible combinations of dataset characteristics are conducted on 128 publicly available datasets. We apply the decision tree method to evaluate the interdependencies between dataset characteristics and performance. The results of the study reveal the intrinsic relationship between dataset characteristics and feature selection techniques' performance. Additionally, our study contributes to research in data mining by providing a roadmap for future research on feature selection and a significantly wider framework for comparative analysis.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

In machine learning, feature selection is the process of identifying and removing irrelevant and redundant features. Feature selection is extremely important in the era of big data to address the large number of input features. Many researchers are focused on developing feature selection techniques and new techniques are constantly introduced. With such a vast body of feature selection techniques, the need arises to determine which technique to use in a given situation. According to the "no free lunch" theory, no best technique exists for all situations [24]. It is therefore imperative to selectively employ appropriate techniques by preprocessing the data. Existing approaches generally use a "trial and error" basis and there is a lack of systematic research concerning which feature selection technique should be used on a particular dataset, based on the characteristics of this dataset. Previous research in the domain of classification techniques suggests that dataset characteristics considerably impact the performance of the classification methods and proves that the choice of the "best" classification algorithm is dependent on the given dataset [5,26,15,4,1,7,29,18,8]. Little recent research has focused on these issues. Kiang suggests that data characteristics considerably impact the performance of classification methods [15]. Bernadoí –Mansilla and Ho [4] have developed metrics to evaluate the ability of a classification problem to characterize various datasets, and use those metrics to determine the appropriate algorithm for a new classification problem. Ali and Smith's study on the classification algorithm selection problem [1] confirms that "a more useful strategy is to gain an understanding of the dataset characteristics that enable different learning algorithms to perform well, and to use this knowledge to assist learning algorithm selection based on the characteristics of the dataset." Chen and Shyu claim that the correlation between the characteristics of a dataset affect the algorithm performance [7], but there are very few studies that analyze the influence of dataset characteristics on classification algorithm performance. The selection of a proper algorithm for a specific classification problem is very difficult, as the choice of the algorithm(s) to use depends on the chosen dataset(s) [29]. Kwon and Sim examine which characteristics of a dataset influence the performance of classification algorithms and prove that classification algorithms show different performances on different data structure types [18]. This brief review of the literature reveals that no single algorithm can perform uniformly well over all datasets. All these studies try to capture the relationship between observed dataset characteristics and classification algorithm performance. There exists some intrinsic relationship

* Corresponding author.
E-mail addresses: dijana.oreski@foi.hr (D. Oreski), stjepan.oreski@kaba.hr (S. Oreski), bozidar.klicek@foi.hr (B. Klicek).

http://dx.doi.org/10.1016/j.asoc.2016.12.023
1568-4946/© 2016 Elsevier B.V. All rights reserved.

between classification algorithm performance and dataset characteristics. If this relationship holds for classification algorithms, we have reason to believe it should also hold for feature selection algorithms. As far as we know, there are no studies that analyze the influence of dataset characteristics on the performance of feature selection techniques. Thus, we propose to address this lack by investigating how dataset characteristics affect the performance of feature selection techniques. Specifically, we consider both the accuracy and time complexity as measures of its performance. We propose to examine experimentally how a dataset's characteristics affect the classifier accuracy and time complexity of various feature selection techniques. We use decision trees to determine the relationships between the techniques, using the dataset characteristics as independent variables and performance metrics as dependent variables.

The paper is structured as follows. Dataset characteristics are described in section 2. Section 3 details the dataset characterization process, feature selection and classification techniques, and the evaluation process. In section 4, we provide research results. Finally, we conclude the paper and give suggestions for further research.

2. Dataset characteristics

Classification is frequently studied by data mining and machine learning researchers [25] and it is considered as a most interesting machine learning challenge [20]. In our research, we build a classification model for strictly binary data, which means a dependent variable is restricted to two possible values. Today, many classifiers are employed in numerous applications. No classifier can outperform all other classifiers on all classification tasks. The first proof that understanding the relationship between dataset characteristics and the performance of classifiers is crucial to the process of classifier selection is from Michie et al. [22]. Furthermore, Sohn claims: 'performance of each algorithm would be closely related to the characteristics of the data to be mined' [28], and Kiang: "data characteristics considerably impact the classification performance of the methods" [15], confirmed by Tax and Duin, [31], and Dessi and Pes [8]. Several empirical studies have shown that the choice of optimal classifier does in fact depend on the dataset employed [22]. Therefore, understanding the relationship between dataset characteristics and the performance of classifiers is crucial to the process of classifier selection. Based on those insights, we hypothesize that the selection of an optimal feature selection technique is determined by dataset characteristics. We want to identify dataset characteristics that influence feature selection techniquesí performance.

Dataset characteristics are grouped into the following categories: standard measures, data sparsity measures, statistical measures, information theoretic measures and noise measures (Table 1). From the 11 characteristics in the Table 1 we have used 7. Number of classes, data sparsity, ID and ID ratio were left out. We deal with binary classification and there are only two classes so number of classes is unnecessary. Feature noise is derived from ID ratio, and ID ratio is derived from ID, thus it was left out.

In the addition we briefly introduce dataset characteristics.

Standard measures are generic measures that serve to normalise many of other measures. Those measures are: number of features, number of instances and number of classes. Number of instances plays a critical role in the selection of classifiers. The number of instances influences classification performance to a great extent, since it determines the amount of information available for the purposes of training [23]. *Data sparsity measures* define relationship between the dimensionality of data and the number of instances required to model the data accurately. This measures

Table 1
Dataset characteristics.

Dataset characteristic	*Acronym*	*Sources*
Standard measures		[22,28,29]
Number of features (Dimensionality)	*d*	
Number of instances	*N*	
Number of classes	*C*	
Data sparsity meaures		[2]
Data sparsity ratio	*DSR*	
Data sparsity	*DS*	
Statistical measures		[22,28]
Correlation of features	*p*	
Multivariate normality	*MVN*	
Homogeneity of class covariances	*SDR*	
Information theoretic measures		[28]
Intrinsic dimensionality	*ID*	
Intrinsic dimensionality ratio	*IDR*	
Noise measures		[19]
Feature noise	*ID2*	

indicate how sparse data is by taking the dimensionality, number of classes and number of instances in a dataset into account [33]. Number of instances that are sufficient to model the data accurately is quantified by defining a ratio between the actual number of instances and the minimum number of instances that are required. *Statistical measures* measure the correlation between features, the multivariate normality of class −conditional probability density functions and the homogeneity of class covariance matrices. We will use Pearson correlation coefficient to measure correlation. The geometric mean ratio between the pooled covariance matrix and the individual class covariance matrices can be used to evaluate the homogeneity of class covariance matrices. The individual class matrices can be tested for homogeneity by making use of Box's M test statistic [for description see 3]. *Information theoretic measures* include intrinsic dimensionality (ID) and ID ratio. The mutual information between classes and features, $M(C;X)$, can be used to determine the intrinsic dimensionality of a dataset. We will measure how many features are not contributing significantly to classification by measuring the importance of features with their values of $M(C;X)$. Intrinsic dimensionality is defined as the number of features required to represent 90% of the mutual information between class and features. We denote this measure as *ID* and the ratio between *ID* and the true dimensionality as *IDR*. If dataset contains a large proportion of features that didn't contribute to classification, dataset has large *feature noise* [19]. The intrinsic dimensionality measure proposed earlier can be used to measure the proportion of features that don't contribute to classification. Feature noise is measured as ratio of difference between dimensionality and ID ratio with dimensionality.

3. Experimental design

Our proposed research is as follows:

(1) Acquire datasets and identify the datasets' characteristics,
(2) Perform feature selection by applying five feature selection techniques to each of the datasets,
(3) Perform classification,
(4) Evaluate the results.

To evaluate the performance and effectiveness of the feature selection techniques on datasets with specific characteristics, verify whether any discovered patterns are potentially useful in practice, and allow other researchers to confirm our results, we detail every step of our experimental study. The experiments described here are performed on a Windows 7 PC with a 2,3 GHz Intel Core i5-4200U

Download English Version:

https://daneshyari.com/en/article/4963453

Download Persian Version:

https://daneshyari.com/article/4963453

Daneshyari.com